Assessing Environmental Impacts of Shale-gas Development in an Area of Hydraulic Fracturing

Guanjie Zheng[†], Fei Wu[†], Matthew Gonzales[§], Susan L. Brantley[§], Thomas Lauvaux[§], Zhenhui Li[†]

[†]College of Information Sciences and Technology, [§]College of Earth and Mineral Sciences
Pennsylvania State University, University Park, PA 16802, USA
{gjz5038,fxw133}@ist.psu.edu{msg5223,sxb7,tul5}@psu.edu,jessieli@ist.psu.edu

ABSTRACT

In this paper, we address an important problem of understanding the impacts of shale-gas development on the groundwater in the state of Pennsylvania, U.S.A., with a focus on the methane leakage from shale-gas wells. The problem is highly challenging because methane concentrations in water are influenced by a large number of factors, many of which are unobserved. Further, the correlation between the methane values and the impacting factors exhibits strong spatial heterogeneity. As a result, simply applying existing data analysis tools would result in poor performance in terms of predicting the methane concentrations values using the available data of the impacting factors, and more importantly, provide little insight into the actual impacts of shale-gas development. In this paper, we take a different approach to develop a simple spatially local ensemble model which is able to explore the local correlations in the data while enabling domain experts to provide timely feedback on its performance. Through experiments on real environmental datasets from multiple sources, we demonstrate the effectiveness of the proposed method.

1 INTRODUCTION

Improvements in horizontal drilling and hydrofracturing has allowed the shale gas extraction, and therefore changed the energy industry. Only 1% of U.S. natural gas production was from shale gas in year 2000, but this ratio might approach 46% by 2035[23]. Figure 1 shows unconventional wells in Pennsylvania before 2008 and as of August 2015. However, the increasing prevalence of shale-gas wells (also called unconventional wells) has potential impacts for contamination of water quality and air quality [21], of which the biggest concern is methane leakage. Methane has the potential to not only cause the contamination of homeowner wells and aquifers [5, 20], but to enhance global warming [4, 16] as well.

Methane leakage from shale gas wells remains notoriously hard to assess. As we can see from Figure 2, there is no clear pattern showing that the number of shale gas wells correlate with methane values in the groundwater. This is because many factors contribute to methane in water, including geology, agriculture, industry, and other anthropogenic activities. For example, many

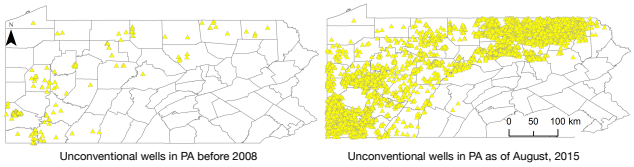


Figure 1: Number of unconventional wells increase from 2008 to 2015 in Pennsylvania (PA), U.S.A.

locations in Pennsylvania show naturally high methane concentration in groundwater due to emissions from the underlying geological formations, due to wetlands, glacial deposits, or due to farm activities.

To systematically assess the environmental impacts from shale gas development, it is critical to study all the potential impacting factors simultaneously and examine their combined impacts. Since the impacting factors are from multiple heterogeneous data sources and some are at a large scale, environmental scientists seek advanced computational tools to help them model such complicated correlations. That means we need to build a robust inference model that can estimate methane concentration by considering all the impacting factors. If shale gas wells play a significant role in such an inference model, it is more likely that the higher methane concentration is caused by shale gas wells (even though we still can not make conclusion about causation).

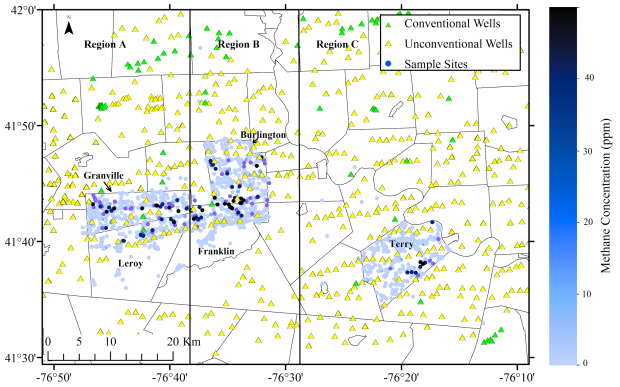
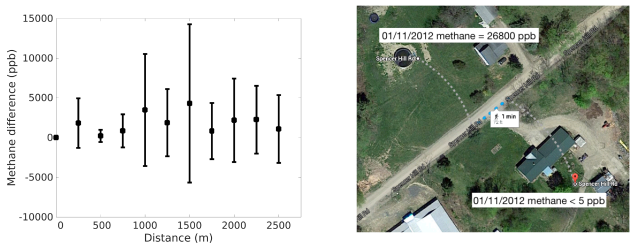


Figure 2: Methane concentrations in groundwater samples [17]

However, the inference of methane concentrations in water, is very challenging due to the following factors. *First, methane concentrations in groundwater have a high spatial heterogeneity.* The

KDD'17 DSIFEW Workshop, Halifax, Nova Scotia, Canada
 978-x-xxxx-xxxx-x/YY/MM…$15.00
DOI: 10.1145/nnnnnnn.nnnnnnn



(a) difference in methane from water samples (b) two sample locations

Figure 3: High spatial heterogeneity for methane concentrations in groundwater.

methane concentrations in groundwater in the northeastern Pennsylvania vary from less than 1 ppb to 46,500 ppb. From Figure 3(a), we see that the average methane difference for data samples within 500 meters is as large as 2,000 ppb. Figure 3(b) shows that, for two water samples collected on the same day with only 100 meter distance, the difference in methane values could be as large as 26, 800 ppb. *Second, the water quality depends on multiple factors, and many of these factors are not fully measured.* For example, fertilizer used on farmland can easily penetrate into groundwater and change the methane concentration but the exact use level and type of the fertilizer can not be recorded accurately.

Therefore, an ideal approach to assess such impacts should (1) be able to exploit the correlation between methane concentration and limited observations of potential factors in the presence of strong spatial heterogeneity, and (2) keep domain experts involved in data interpretation so that they can provide timely feedback as well as additional data to improve the model performance. In the data mining literature, numerous models have been proposed which can handle spatial heterogeneity with varying degrees of success. However, little attention has been paid to the inclusion of domain experts. As a result, even if the existing models are able to achieve decent accuracies, the results are often very hard to interpret.

In view of this difficulty, the key innovation of this paper is *a spatially local ensemble model* which is able to explore local correlations, and provide our collaborators in geoscience and meteorology with new insight to the problem. On one hand, we observe that correlation between methane and factors could vary at different locations. For example, at certain regions, the methane concentration has a strong correlation with number of unconventional wells, whereas in other regions, such a correlation is very weak. Therefore, building a local inference model enables us to make more accurate inference. On the other hand, such a local model allows geoscientists to reason about results on a map and take further actions when necessary. For example, in our study [17], we visualized the local importance of unconventional wells, and geoscientists were able to infer that such a correlation could be due to a geological factor - faults. Fault data were subsequently collected. Even though the data are incomplete, this factor could be further considered in the inference model.

The contribution of this paper lies in the following aspects:

- We propose to adopt a spatially local ensemble learning approach, which effectively uses the local correlations for inference. We evaluate our method using data from multiple sources including water quality data, gas wells data, industry emission data, and land use data.
- We collaborate with geoscientists and meteorologists in analyzing our models. As an example of the interactive data exploration process, we conduct in-depth studies to understand the model in terms of model uncertainties and outliers. These explanations help domain experts to further collect more samples to understand the correlations and outliers. This also helps us to build more systematic models to analyze the water quality data.

The rest of the paper is organized as follows. We first review the related work in Section 2. Then we introduce the datasets and describe how to construct features in Section 3. We describe our correlation discovery and inference model in Section 4. Experimental results are presented in Section 5. And finally we conclude the paper in Section 6.

2 RELATED WORK

Generative models. In the field of environmental science and geoscience, generative models [24] are frequently used for environmental reading prediction. These models often incorporate meteorology, street geometry, receptor locations, traffic volumes, and emission factors, based on some empirical assumptions. Such assumptions might not be applicable to all scenarios [29]. At the same time, these parameters and data are difficult to obtain precisely, thus the results generated by these methods may not be very accurate [28, 29]. Therefore, we propose to use a data-driven discriminative model to learn the correlations between water quality and available features.
Discriminative models. Machine learning and statistical models, such as linear regression [9], classification and regression trees (CART) [25], neural network [13, 18, 22, 25], and support vector machine [13], have been adopted by geoscientists to predict air quality [9, 25, 26] and water quality [13, 18]. Recently, Zheng et al [28, 29] discussed how to use advanced data mining techniques to predict PM2.5 and PM10 in the air. They propose a co-training method to infer air quality data [28], and use a regression tree [29] to predict air quality. Our work is different from these work in three aspects. First, we observe local correlations and use a local ensemble model to describe such local correlations. Previous methods have not considered such spatial heterogeneity at the model level. Second, the data used in previous studies are often collected by sensors at fixed locations in the format of a time series. Therefore, historical temporal information can be used for prediction. This is not the case for our problem because our data are collected at different locations and temporally sparse. Third, most previous studies use environmental quality data of nearby locations to infer or predict the environmental quality at a given location. We do not use any environmental quality data as input to our model because our main goal is to understand the correlation between the impacting factors and environmental quality data instead of making accurate predictions.

3 DATASETS AND FEATURE CONSTRUCTION

In this section, we provide details about our datasets and features used in our study.

3.1 Water Quality Data

We study 1,816 water quality samples, sampled from Year 2010 to 2015. These water samples were taken for point locations and there are no data available for other locations. Each data sample has values for 30 to 50 chemical analytes. In this paper, we focus on methane values because methane is the major contamination concern from shale-gas development. Among these data samples, 1,684 of them have methane concentration values. This dataset can be found at [3]. While thousands of data samples may look "small", this is the largest public groundwater data because it is expensive to sample and measure the chemical analytes in groundwater.

As illustrated in Figure 2 and Figure 3, methane concentrations in groundwater have high spatial heterogeneity. The values could be very different for two spatially close locations. Also, the distribution of methane values are highly skewed. As we can see from the histogram in Figure 4, there are many samples with values 5 ppb and 26 ppb. These two numbers are the detection limits, where the actual values are not reported once the values are below these limits. We observe 71% data points have values lower or equal to 26 ppb but the large values could be as high as 46,500 ppb.

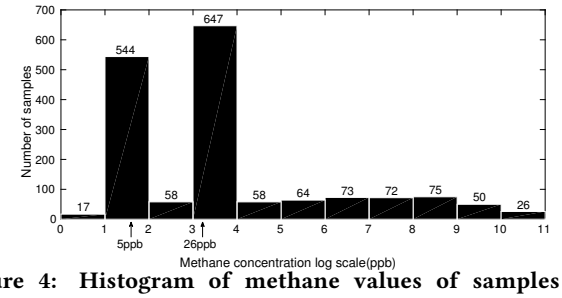


Figure 4: Histogram of methane values of samples in groundwater.

3.2 Feature: Wells

Wells data include both conventional wells and unconventional wells. A conventional gas well produces oil or gas from a conventional formation while an unconventional well usually employs sophisticated methodologies including horizontal drilling and hydraulic fracturing. In our studying area, as of May, 2015, there are 61,739 conventional wells and 6,227 unconventional wells. The well data are obtained from Pennsylvania Department of Environmental Protection (PA DEP) [2].

Figure 2(a) shows the conventional wells and unconventional wells in our focused area as of May, 2015. In our work, we use the 2,276 conventional wells and 3,279 unconventional wells located in the northeastern Pennsylvania, where the water quality samples are located.

Well emission data estimate how much methane (in mol/hr) are let into the air due to the production of gas wells. It is derived from well production data by assuming that there is a leakage rate of 0.2% for unconventional well and 5% for conventional well. Figure 5(a) and (b) show the estimated emission of methane from conventional wells and unconventional wells, respectively. Based on locations and emissions from both conventional wells and unconventional wells, we construct features as shown in Table 1.

Similar to that in [17] (different factors are used), figure 6 plots the methane values in water w.r.t. the feature well_emission_total from unconventional wells with a distance threshold of 5 kilometers. In fact, we do not observe a significant correlation between methane concentrations and the emissions from unconventional wells. But this does not mean that such a correlation does not exist in certain regions. We have studied the local correlations in details in [17].

3.3 Feature: Industrial Emission

Industrial emission data is collected from Greenhouse Gas Protocol [1] by the Greenhouse Gas Reporting Program in 2014. There are 7,289 reported industrial facilities in total, of which 331 are located in Pennsylvania. Each facility will report their emission of methane, carbon dioxide, nitrous oxide and several other emissions. Figure 5(c) displays the industry emission of methane on the map. Similar to the wells, we construct five features using industrial emission data as shown in Table 1.

3.4 Feature: Land Use

Land use data is collected from National Land Cover Database [15]. It reflects a combination of human activities and natural geology. In this study, we use land use image with 30-meter resolution (the highest resolution we could obtain). In Figure 7, we show the land use types of water sample locations.

Figure 8 shows the distribution of methane values for different land types. The methane values (water) are generally higher in wetlands, as expected.

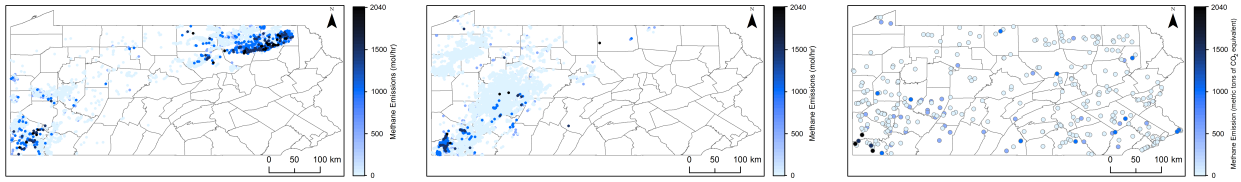
3.5 Feature: Geology

We also consider a set of geological features in our study, including elevation and distance to fault (see Table 1).

4 CORRELATION DISCOVERY AND INFERENCE MODEL

In this section, we develop inference models to predict methane concentrations in water using the aforementioned features. Note that our feature set does not include methane concentrations from nearby locations or historical data. This is in strong contrast to the inference model on air quality in [28, 29], whose ultimate objective is to make accurate predictions using both historical air quality data and contextual data. The reason is that we wish to discover the correlations between the features and methane concentrations, and use such correlations to help domain experts interpret the observed methane concentrations. If we include methane concentrations from nearby locations as features, the inference model is likely to be dominated by facts such as "the methane concentration is likely to be high if nearby concentrations are high". Such facts are not helpful in mining the root cause of high methane concentrations at certain locations.

A baseline method is to build a global classification model using all the training samples. However, such method is problematic as it fails to account for local correlations due to spatial heterogeneity. As we discussed in previous sections, a feature could have very different correlations with the methane concentration at different locations. For example, the hydrofracturing could have a much



(a) Unconventional well emission (b) Conventional well emission (c) Industrial emission

Figure 5: Methane emission from conventional wells, unconventional wells, and industry.

Table 1: Features used in our study.

Feature Name	Descriptions
Wells:	
dist_to_conv_well/dist_to_unconv_well	distance from sample location to closest conventional/unconventional well.
conv_well_num/unconv_well_num	number of conventional/unconventional wells within certain distance threshold.
conv_well_density/unconv_well_density	density of conventional wells/unconventional wells.
conv_well_emission_total/unconv_well_emission_total	total methane emission by conventional wells/unconventional wells within certain distance threshold.
conv_well_emission_density/unconv_well_emission_density	density of methane emission by conventional wells/unconventional wells.
Industry:	
dist_to_industry	distance from sample location to the closest industrial facility.
industry_num	number of industrial facilities within certain distance threshold.
industry_density	density of industrial facilities.
industry_emission_total	total methane emission by industrial facilities within certain distance threshold.
industry_emission_density	density of methane emission by industry facilities.
Land Use:	
land_use_type	type of land use.
Geology:	
elevation	common geological features.
distance_to_fault	distance from sample location to fault.

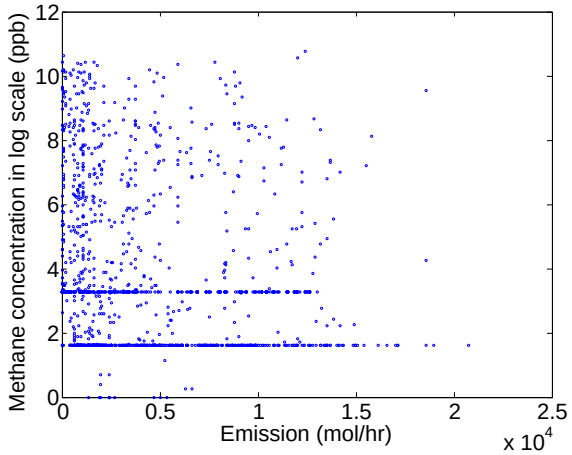


Figure 6: Methane concentration w.r.t. emission from unconventional well.

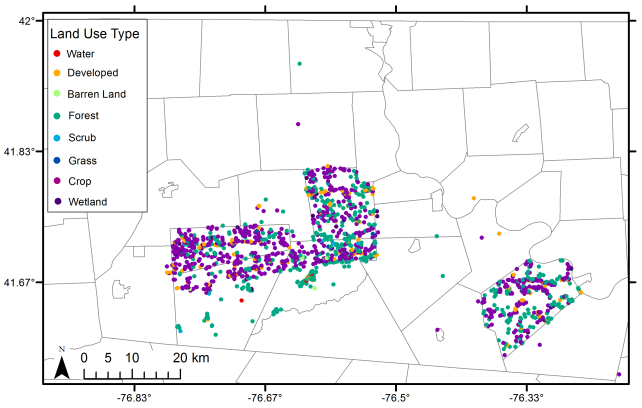


Figure 7: Land use distribution.

larger impact on water quality if the wells are drilled close to faults [17].

Figure 9 shows our proposed framework which explores the local correlations while having domain experts in the loop. The discovery process starts with observing the local correlations. Statistical analyses are first applied to regions to reveal local correlations in Section 4.1. Next, we develop machine learning methods to train a local inference model in Section 4.2. Since domain experts care more about the interpretability of the model than the prediction accuracy, we discuss various approaches to interpret the model in Section 4.3.

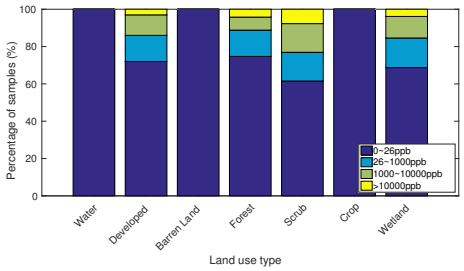


Figure 8: Methane concentration w.r.t. land type.

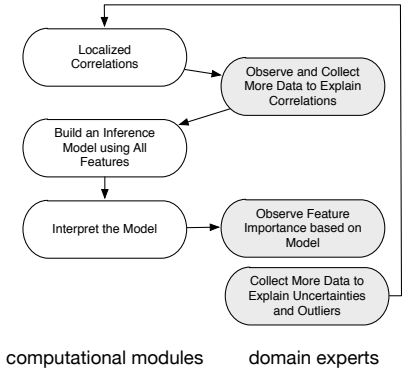


Figure 9: The flow of discovery process.

4.1 Observing Local Correlations

We first study the correlation between individual features and methane concentrations in the water. This correlation analysis is often conducted as a first step because people may intuitively raise the question such as "does the development of shale gas wells increase the methane in the water?" To answer these questions, essentially we wish to know if there is a correlation between features related to unconventional wells and the methane concentration.

Kendall Rank Correlation. Many metrics could be used to measure the correlation between a feature and methane values, such as linear regression coefficient, Pearson correlation, and Kendall rank correlation. Here we choose Kendall rank correlation because there are censored values in our dataset as discussed in Section 3 and Kendall rank correlation can deal with environmental data with multiple detection limits [14]. Kendall rank correlation is calculated as:

$$\tau = (n_c - n_d)/n_0,$$

where n_c is the number of concordant pairs, n_d is the number of discordant pairs, and $n_0 = \frac{n(n-1)}{2}$ with n being the number of sample points in the dataset.

Local Correlation. Given the two variables (e.g., methane concentration and distance to unconventional well), we use Kendall rank correlation to measure whether they are positively or negatively correlated and whether such a correlation is significant. Such a correlation could vary at different regions. We have shown in [17] that the correlation values for different regions vary a lot.

4.2 Spatially Local Ensemble Model

Ensembles are well-known as a method for obtaining high accuracy classifiers by combining less accurate ones [6, 7, 11, 12]. Let $\{y_i\}_{i=1}^n$ be a set of responses each associated with p predictors

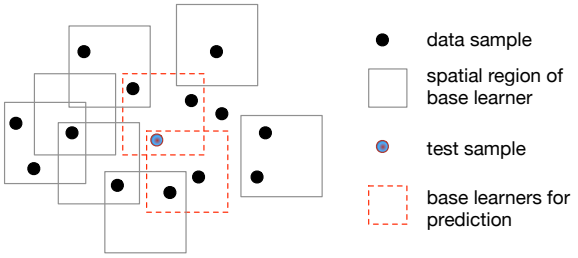


Figure 10: Illustration of the ensemble model.

$\mathbf{x}_i = [x_{i,1}, x_{i,2}, \ldots, x_{i,p}]$. An ensemble model $F(\mathbf{x})$ makes prediction based on a set of base learners $\{f_m(\mathbf{x})\}_{m=1}^M$:

$$F(\mathbf{x}) = \sum_{m=1}^{M} \alpha_m f_m(\mathbf{x}).$$

Here, each $f_m(\mathbf{x})$ is a different function of the predictors $\mathbf{x}$ learned from the data, an α_m defines the weight of $f_m(\mathbf{x})$.

In our problem setting, we adopt a local ensemble model [10] which restricts each base model to a local spatial window. Each base model f_j is trained using data samples in a spatial window S_j. To make the prediction for a data sample $\mathbf{x}$ at location s, we use all the base models f_j where the corresponding spatial window S_j covers this location s:

$$F(\mathbf{x}(s)) = \frac{1}{Z} \sum_j f_j(\mathbf{x}(s)) I(s \in S_j).$$

$I(s \in S_i)$ is the indicator function, taking value 1 when the location is within window S_i, and 0 otherwise. In addition, Z calculates the number of base learners supporting the prediction at location s:

$$Z = \sum_j I(s \in S_j).$$

For simplicity, all the base learners have equal weights.

Figure 10 illustrates the idea of our spatially local ensemble model. Each ensemble prediction is made for a particular location and is computed as the average prediction (or majority vote) made by all base models that contain that location. The behavior of such a local prediction is similar to that of parametric models for spatial and temporal correlation giving larger weights to nearby observations.

Next, we discuss how to choose the base learners and how to select local spatial window S_j to train the base learners.

4.2.1 Decision Tree as the Base Learner. The base learner can be any supervised learning model. In our problem setting, we choose the decision tree because it can handle non-linear correlations easily. Using decision tree as the base learner, our approach is similar to that of a random forest [7], in which we build many decision trees, each using a subset of training samples, and combine their outputs to obtain the final result. Random forests are based on the idea of bagging, and are known to produce very good predictive accuracy [8, 19] because bagging can effectively reduce the variance in the model. The difference in our method is that, instead of randomly sampling the training data, we use the subset of training samples that are spatially close (e.g., in a region) to learn the base models. In this way, the base models will preserve the local correlations.

4.2.2 Density-Based Spatial Windows. There are two ways to generate the spatial windows to build an ensemble model. One way is to uniformly sample the locations on the map. However, by doing so, some base learners might only cover very few samples.

An alternative strategy is to uniformly sample the training data and use each sample's location as the center of the window. In this way, we will generate more base models for locations with higher density of observations. In our experiment, we found that the second approach yields a better performance.

4.3 Interpreting the Model

The ensemble models (e.g., random forest) often have better predictive accuracy, but they are often considered as "black-box" models, which are hard to interpret directly. Fortunately, there are various heuristics we can use to "probe" such models, in order to answer following questions: (1) What are the locations that the model is uncertain in making the prediction? (2) Are there any outliers in the data?

The spatially ensembled model often have different confidence levels in making predictions at different locations, especially at the locations that are more heterogeneous in geology or have less training data. By taking the advantage of our proposed local model, we can further visualize such confidence scores on a map. For a location, we consider all the decision trees covering this location. If the predictions made by these trees have a big discrepancy, a low confidence in prediction is indicated. On the other hand, if all the trees have little discrepancy (i.e., high confidence), but the predicted value is very different from the true value, this data sample is likely to be an outlier. In both cases, the corresponding data samples are worth further investigation (e.g., finding more factors for explanation or collecting more samples).

5 EXPERIMENTAL EVALUATION

5.1 Experiment Settings

Baseline Methods. We compare our method with five classification models including decision tree (DT), random forest (RF), Logistic regression (LR), support vector machine (SVM), and artificial neural network (ANN). ANN is used in a recent study by Zheng et al. [28], which attempts to classify the levels of PM2.5 and PM10 in the air quality data. Note that the paper [28] proposes to use both a temporal classifier and a spatial classifier. While their air quality data are collected at several fixed monitor stations [28], our data are collected at different spatial locations and there is no continuous data for multiple timestamps at the same location. Due to this difference in data collection, we only compare with the spatial classifier in [28], which is the ANN model.

Ground truth and Evaluation. For water quality data, we choose medium value (26ppb) as the threshold and classify whether methane values are above the medium or not (i.e., binary classification).

Parameter settings. For our local ensemble model, we sample 500 square windows and train a decision tree for each window. We set the minimum number of data samples for each tree to be 50 (any tree with less than 50 samples will be discarded and a new one will be selected). We use windows with multiple sizes (5,000m, 10,000m, and 100,000m) .

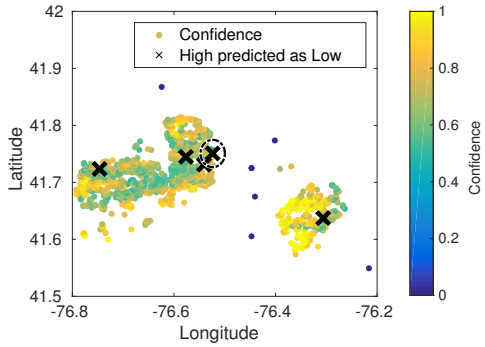
5.2 Experiment Results

Table 2 compares the performance of our local ensemble model with the five baselines. As one can see, our model outperforms other methods in terms of the accuracy, which shows that the local correlations indeed play an important role in predicting the methane concentration value. In terms of the standard deviation, none of these methods has a standard deviation larger than 0.01, indicating that their performance is very stable.

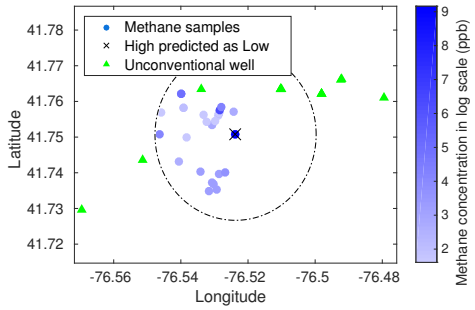
Table 2: Prediction accuracy of all methods. Standard deviations are given in the parentheses.

LR	0.716 (3.0e-3)
SVM(linear)	0.723 (3.5e-3)
SVM(rbf)	0.723 (2.8e-4)
ANN(mlpc)	0.697 (1.5e-3)
DT	0.687 (3.4e-3)
RF	0.718 (4.7e-3)
Local Ensemble	**0.731 (5.8e-3)**

5.3 Model Interpretation



(a) Confidence map and outliers



(b) Zoomed-in map of one outlier

Figure 11: Confidence map and outliers.

As discussed in Section 4.3, we define prediction confidence of one test sample as the number of votes to a majority vote label divided by the total number of votes covering that sample. Figure 11(a) shows the prediction confidence of our method on all water samples. (This confidence map might look slightly different for different random runs of experiments.) One can clearly see that our model performs better in certain regions than the others. Such observations are of potential value to our collaborators in

geoscience as they reveal problematic regions on which further study should be focused.

Using the confidence map, we are also able to identify outlying samples in the dataset. The black crosses in Figure 11(a) indicate five such samples, which have methane concentration values higher than 5,000 ppb but are predicted as lower than 26 ppb, each with a confidence value higher than 0.8. We plot the nearby region on the map in Figure 11(b). There are 30 samples falling into this circle of 2km in total. We found that the distance to unconventional wells is a discriminative feature for this region. When the distance is smaller than 900m, there are 5 low methane concentration samples and 5 high methane concentration samples, while when the distance is more than 900m, there are 17 low methane concentration samples and 3 high methane concentration samples (see Figure 11(b)). Since the focal sample is very far from any unconventional well, it is predicted to have a methane concentration lower than 26 ppb by our model with a high confidence. However, it actually has a high methane value that makes it as an outlier. To further explain such outliers, we will work with our collaborators in geoscience to obtain more water data samples around that location and to make some field trips to learn more about the geographical structure or other anthropogenic activities near that location. Such data could be further incorporated into our model in the next round of analysis. With the help of our domain experts in analyzing the results, we have successfully delivered another more systematic spatial outlier detection algorithm in [27], which has been proved to be able to discover potential leakage problems in the ground water dataset.

6 CONCLUSION

In this paper, we propose to use a spatially local ensemble model to assess the influence of shale gas development on groundwater quality. Multiple factors are considered in our study, including gas wells, industrial emission, land use, and geological features. We have demonstrated that our local ensemble model outperform the global models in predicting the methane concentrations. We have further interpreted our model by studying the prediction confidence and detecting outliers. Such interpretations enable domain experts to identify specific areas for further investigation.

ACKNOWLEDGMENTS

The work was funded by a gift to Penn State for the Pennsylvania State University General Electric Fund for the Center for Collaborative Research on Intelligent Natural Gas Supply Systems and was supported in part by NSF awards #1639150, #1618448, #1652525, and #1544455. This work has also been funded by the U.S. Department of Energy National Energy Technology Laboratory (project DE-744FE0013590). We also thank Anna K. Wendt from College of Earth and Mineral Sciences, Penn State University for her contribution in this work. The conclusions and observations in this paper are from the authors and do not represent the view of any funding agencies.

REFERENCES

[1] 2014. Greenhouse Gas Protocol. http://www.ghgprotocol.org/. (2014).
[2] 2015. PA DEP Oil & Gas Reporting Website. (2015). https://www.paoilandgasreporting.state.pa.us/publicreports/Modules/Welcome/ProdWasteReports.aspx
[3] 2015. Shale Network. (2015). https://doi.org/10.4211/his-data-shalenetwork
[4] Ramón A Alvarez, Stephen W Pacala, James J Winebrake, William L Chameides, and Steven P Hamburg. 2012. Greater focus needed on methane leakage from natural gas infrastructure. *Proceedings of the National Academy of Sciences* 109, 17 (2012), 6435–6440.
[5] SL Brantley, D Yoxtheimer, S Arjmand, P Grieve, R Vidic, J Pollak, GT Llewellyn, JD Abad, and C Simon. 2014, doi: 10.1016/j.coal.2013.12.017. Water resource impacts during unconventional shale gas development: tfhe Pennsylvania experience. *International Journal of Coal Geology* (2014, doi: 10.1016/j.coal.2013.12.017).
[6] Leo Breiman. 1996. Bagging predictors. *Machine learning* 24, 2 (1996), 123–140.
[7] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.
[8] Rich Caruana and Alexandru Niculescu-Mizil. 2006. An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning*. ACM, 161–168.
[9] Aoife Donnelly, Bruce Misstear, and Brian Broderick. 2015. Real time air quality forecasting using integrated parametric and non-parametric regression techniques. *Atmospheric Environment* 103 (2015), 53–65.
[10] Daniel Fink, Wesley M Hochachka, Benjamin Zuckerberg, David W Winkler, Ben Shaby, M Arthur Munson, Giles Hooker, Mirek Riedewald, Daniel Sheldon, and Steve Kelling. 2010. Spatiotemporal exploratory models for broad-scale survey data. *Ecological Applications* 20, 8 (2010), 2131–2147.
[11] Yoav Freund and Robert E Schapire. 1996. Experiments with a new boosting algorithm. In *ICML*, Vol. 96. 148–156.
[12] Jerome H Friedman and Bogdan E Popescu. 2008. Predictive learning via rule ensembles. *The Annals of Applied Statistics* (2008), 916–954.
[13] Zhibin He, Xiaohu Wen, Hu Liu, and Jun Du. 2014. A comparative study of artificial neural network, adaptive neuro fuzzy inference system and support vector machine for forecasting river flow in the semiarid mountain region. *Journal of Hydrology* 509 (2014), 379–386.
[14] Dennis R Helsel. 2011. *Statistics for censored environmental data using Minitab and R*. Vol. 77. John Wiley & Sons.
[15] C.G. Homer, J.A. Dewitz, L. Yang, S. Jin, P. Danielson, G. Xian, J. Coulston, N.D. Herold, J.D. Wickham, and K. Megown. 2015. Completion of the 2011 National Land Cover Database for the conterminous United States-Representing a decade of land cover change information. *Photogrammetric Engineering and Remote Sensing* 81, 5 (2015), 345–354.
[16] Robert W Howarth, Renee Santoro, and Anthony Ingraffea. 2011. Methane and the greenhouse-gas footprint of natural gas from shale formations. *Climatic Change* 106, 4 (2011), 679–690.
[17] Zhenhui Li, Cheng You, Matthew Gonzales, Anna K Wendt, Fei Wu, and Susan L Brantley. 2016. Searching for anomalous methane in shallow groundwater near shale gas wells. *Journal of Contaminant Hydrology* 195 (2016), 23–30.
[18] Saumen Maiti and RK Tiwari. 2014. A comparative study of artificial neural networks, Bayesian neural networks and adaptive neuro-fuzzy inference system in groundwater level prediction. *Environmental earth sciences* 71, 7 (2014), 3147–3160.
[19] Kevin P Murphy. 2012. *Machine learning: a probabilistic perspective*. MIT press.
[20] Stephen G Osborn, Avner Vengosh, Nathaniel R Warner, and Robert B Jackson. 2011. Methane contamination of drinking water accompanying gas-well drilling and hydraulic fracturing. *proceedings of the National Academy of Sciences* 108, 20 (2011), 8172–8176.
[21] Michael Ratner and Mary Tiemann. 2014. An overview of unconventional oil and natural gas: resources and federal actions. *Congressional Research Service Report* 43148 (2014).
[22] Silvia L Reich, DR Gomez, and LE Dawidowski. 1999. Artificial neural network for the identification of unknown air pollution sources. *Atmospheric Environment* 33, 18 (1999), 3045–3052.
[23] Paul Stevens. 2012. The fishale gas revolutionfi: Developments and changes. *Chatham House* (2012), 2–3.
[24] Sotiris Vardoulakis, Bernard EA Fisher, Koulis Pericleous, and Norbert Gonzalez-Flesca. 2003. Modelling air quality in street canyons: a review. *Atmospheric environment* 37, 2 (2003), 155–182.
[25] Yang Zhang, Marc Bocquet, Vivien Mallet, Christian Seigneur, and Alexander Baklanov. 2012. Real-time air quality forecasting, part I: History, techniques, and current status. *Atmospheric Environment* 60 (2012), 632–655.
[26] Yang Zhang, Marc Bocquet, Vivien Mallet, Christian Seigneur, and Alexander Baklanov. 2012. Real-time air quality forecasting, part II: State of the science, current research needs, and future prospects. *Atmospheric Environment* 60 (2012), 656–676.
[27] Guanjie Zheng, Susan L. Brantley, Thomas Lauvaux, and Zhenhui Li. 2017. Contextual Spatial Outlier Detection with Metric Learning. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM.
[28] Yu Zheng, Furui Liu, and Hsun-Ping Hsieh. 2013. U-Air: When urban air quality inference meets big data. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1436–1444.
[29] Yu Zheng, Xiuwen Yi, Ming Li, Ruiyuan Li, Zhangqing Shan, Eric Chang, and Tianrui Li. 2015. Forecasting Fine-Grained Air Quality Based on Big Data. In *Proceedings of the 21th ACM SIGKDD international conference on Knowledge discovery and data mining*.